Task Specification: GALE Evaluation Data Selection
Version 2.2, January 2, 2007

1 Goal

The goal of this effort is to select suitable test data for inclusion in the GALE Phase 2 Go/No-Go Translation Evaluation.

Test data is drawn from newswire, broadcasts and web text collected by LDC during the designated evaluation epoch; this collection is known as the *candidate pool*. Humans review the contents of the candidate pool to identify documents or files that meet the selection criteria described below. These manually selected documents constitute the *selection pool*. The selection pool is then subject to automatic scans to flag outliers. At a later stage, the selection pool is subsampled to identify the specific documents and files that will comprise the final *evaluation data set*.

2 Data Profile

2.1 Evaluation Epoch

The evaluation epoch is designated as November 1 through December 22, 2006. Selection is limited to documents/files that are originally published within these dates[1].

2.2 Source Languages and Dialects

Source documents/files will be selected in Modern Standard Arabic and standard Mainland (Beijing dialect) Chinese. In general, documents/files that contain more than the occasional word outside of the target language and dialect will not be selected. However, Arabic web text frequently contains significant portions of Arabic dialects in addition to MSA. For purposes of Arabic web text (newsgroups and weblogs), dialect data will not be excluded from the selection, but will be flagged as containing colloquial Arabic in LDC's selection tracking database.

2.3 Genre

There are four genres: ***newswire*** (NW), ***broadcast news*** (BN), ***broadcast conversation*** (BC), and ***web text*** consisting of a mix of weblogs (WL) and newsgroups (NG).

2.4 Data Volume

The selection pool will comprise approximately 150 documents or files per language per genre. Within each selected document a contiguous region of approximately 150-250 words will be identified (equivalent to roughly 1.5-2.5

[1] The selection of evaluation data may come from later in this epoch, leaving open the possibility of a devtest set to be selected from the early part of the epoch.

minutes for Arabic audio sources, 1-1.5 minutes for Chinese). The resulting selection pool will consist of between 30,000 and 45,000 words per genre. The final evaluation data set drawn from the selection pool will consist of approximately 15,000 words per language per genre.

2.5 Document Difficulty

All documents will be categorized by difficulty level, using Interagency Language Roundtable (ILR) difficulty ratings. LDC annotators will provide ILR ratings during the initial selection process. Although there are no hard and fast requirements regarding document difficulty, the selection process will primarily target level 2 and level 3 documents.

- Level 1 texts: contain short, discrete, simple sentences; generally pertain to the immediate time frame; often written in an orientational mode; require elementary level reading skill. *Example: Newspaper announcements.*
- Level 2 texts: convey facts with the purpose of exchanging information; do not editorialize on the facts; often written in an instructive mode; require limited working proficiency. *Example: Newswire articles; TIDES/MT evaluation data.*
- Level 3 texts: have denser syntax and highly analytic expressions; place greater conceptual demands on the reader; often written in an evaluative mode; may require the reader to 'read between the lines'; require general professional proficiency. *Example: newspaper opinion / editorial articles*.
- Level 4 texts: express creative thinking; assume a relative lack of shared personal information; often involve a highly individualized mode that projects the style of the author; require advanced professional proficiency. *Example: essays; political editorials that reformulate social, economic or political policy.*[2]

2.6 Topic content

An effort will be made during selection of the selection pool to vary topic content within and across genres.

2.6.1 Suitable topic content

Suitable topics for selection include

- hard news, politics and current events - local, national or international
- social issues
- human interest stories

2 Clifford, R., Granoien, N., Jones, D., Shen., W, Weinstein, C. " The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean". http://www.mt-archive.info/LREC-2004-Clifford.pdf

- opinions about current events, news, politics, social issues
- reactions to current events, news, politics, social issues
- editorials about current events, news, politics, social issues
- interviews, roundtable discussions, call-in shows whose focus is current events, news, politics, social issues

2.6.2 Topic content to exclude

Among topics to be excluded are the following:

- re-broadcast or re-print of material originally published from another source
  - e.g., a weblog post that contains a selection of text from a newspaper
  - e.g., a news broadcast that features a clip from another show
- brief headline reviews (multiple headlines covered in a very brief segment)
- previews for upcoming broadcasts
- broadcast of sporting events
- lists of any kind, including lists of URLs, weather statistics, sports scores or stock figures
- weather reports
- recipes and how-to guides
- stock reports
- public service announcements
- commercials, advertisements and classifieds
- infomercials
- music features
- movie or book reviews
- top ten countdowns and the like
- religious texts (prayers, supplications, poetry, songs)
- excerpts from novels, books, poems, songs or other literary genres
- interviews, roundtable discussions, call-in shows whose focus is solely entertainment topics (celebrities, movies, musical acts)
- holiday or year-end feature stories
- embedded links to audio/video files
- jokes (social networking sites are full of these)
- horoscopes and the like
- quizzes, memes
- chain letters
- documents containing “sensitive” material (sexually suggestive or explicit content, frequent use of obscene language, insults)
- racist or “hate" documents
- scam letters (e.g., “I am from Nigeria, please send money”)
- fake news (e.g., “The Onion”)

2.7 Genre-specific Issues

2.7.1 Newswire

No known issues.

2.7.2 Broadcasts

Each show collected for GALE is pre-designated as Broadcast News (BN) or Broadcast Conversation (BC) based on its characteristic content. Note however that BN shows can sometimes contain stories that are conversational, while BC shows can include hard news reports. When selecting files for inclusion in the test set, annotators should identify segments that are typical of their designated genre. BN segments might include talking head news reports, feature stories from field reporters or other segments typical of a news broadcast. BN selections should not include highly interactive, conversational segments (such as an in-depth interview with a guest). BC segments might include interviews, roundtable discussions or other interactive segments. BC selections should not include talking head or single-speaker field reports.

2.7.3 Web data

Web data selections should be limited to newsgroups and weblogs. Newsgroups are repositories or archives for many messages posted by many users from many different locations. For purposes of GALE, the term newsgroup may also include discussion groups or discussion forums, which are technically distinct from but functionally similar to newsgroups.[3] Weblogs (or blogs) are websites made in journal style, displayed in reverse chronological order. Weblogs may be maintained by one person or by a group of people. A weblog entry typically contains the following features:

- *Title*, the main title, or headline, of the post.
- *Body*, main content of the post.
- *Permalink*, the URL of the full, individual article.
- *Post Date*, date and time the post was published.
- A blog entry optionally includes the following:
- *Comments*
- *Categories* (or tags) - subjects that the entry discusses
- *Trackback* and or *pingback* - links to other sites that refer to the entry[4]

Chat rooms, email archives and web text like ezines are not eligible for selection.

Weblogs and newsgroups frequently contain long snippets of text extracted from other sites, including copyrighted material from other news providers. This

[3] Adapted from http://en.wikipedia.org/wiki/Newsgroup

[4] See http://en.wikipedia.org/wiki/Blog for this definition and for more detailed descriptions and examples of weblogs.

material should be excluded from the selection. Our target is material that is posted by the original author.

3 Selection of segments from a larger document or file

Newswire data is presented for manual review as individual story units, with story boundaries pre-defined by the source data provider. Broadcasts are presented for review as single audio files that correspond to a recording of one episode of a single program, typically 20, 30 or 60 minutes in duration. Weblogs and newsgroups are presented for review as threads, which consist of a single entry or post on a given topic, plus all the follow-up entries, posts, messages and/or comments responding to the original post.

The length of candidate documents and files will vary considerably. For each document, annotators must select a smaller contiguous span of text or audio for inclusion in the selection pool. These smaller units are called *segments*. The targeted length of the selected segments is approximately 150-250 words (1.5-2.5 minutes for Arabic, 1-1.5 minutes for Chinese)[5].

Selected segments should comprise a topically contiguous portion of the broadcast, story or thread. Segments must have some kind of natural cohesion, functioning as a snippet of representative content from the larger document or file. They should correspond to some naturally occurring unit within the larger document or file, such as a paragraph from a newswire story, or a single post from within a larger newsgroup thread. See the sections that follow for genre-specific information about what constitutes a natural unit for selection.

Annotators should select no more than one segment from each newswire or web text document. Annotators may select multiple segments from a single broadcast, provided those selections constitute separate stories.

3.1 Newswire

Natural units for newswire documents will typically consist of paragraphs. If a document contains suitable material but the paragraph exceeds 300 words, annotators may select part of a paragraph. However, the selection must consist of full sentences – that is, the selection cannot begin or end in the middle of a sentence. Annotators may select paragraphs from anywhere in a document –the beginning, the middle or the end, provided the paragraph constitutes a natural, topically-contiguous unit. The selection locale should vary across documents and sources so that the final evaluation pool contains material from different parts of newswire documents.

[5] For speech sources, selection is done from the audio recording, not from a transcript.

3.2 Broadcasts

The natural units to be selected will typically correspond to a single story within a broadcast. A story is defined as a topically contiguous segment of the broadcast. In some cases, stories are easy to recognize and correspond directly to a single hard news report, a single interview segment, a single human interest feature, etc. In other cases story boundaries are tricky to recognize. For instance, news stories may discuss more than one related topic. When reports of similar content are adjacent to one another in a broadcast, it is often difficult to tell where one story ends and the next begins. Similarly, conversational stories can be quite long and can contain multiple topics.

Because of this, annotators should use their judgment, and if necessary can select something less than a full interview or full news story to constitute a natural unit to serve as input data. However, selections should be made only at natural breaks in the flow of conversation, for instance, when there is a major shift in topic, or when a new panelist joins a roundtable discussion. Annotators should rely on both the content of the segment (i.e., the topic) as well as audio cues (speaker changes, music, pauses) to inform their judgments.

Annotators should not include two obviously distinct stories in a single selection. If two adjacent stories are to be selected, the annotator should label them as two separate selections rather than merging them together. Annotators should not under any circumstances select commercials, musical interludes, public service announcements or other types of non-news or non-conversational data. If a story spans a commercial break, annotators should either select only one portion of the story, or exclude the story from selection.

3.3 Web Text

Natural units for newsgroup and weblog documents will typically comprise a single post, entry or comment/response within a thread. For long posts/entries, a paragraph within that post may be selected. Annotators may also select a sub-paragraph unit for inclusion when the full paragraph exceeds 300 words. As with newswire sources, the selection must contain full sentences – that is, the selection cannot begin or end in the middle of a sentence. Annotators may select (sub-)paragraphs from anywhere in a document –the beginning, the middle or the end, provided that the paragraph constitutes a natural, topically-contiguous unit. The "selection locale" should vary across documents and sources so that the final evaluation pool contains material from different parts of web documents.

4 Manual Data Selection Toolkit and Methodology

LDC has developed a customized user interface for doing data selection. The interface uses a database backend to track annotation decisions.

The process begins by dividing the collected evaluation data into randomized file lists. Each list includes up to 50 files[6], drawn from a single language and genre (broadcast, web or newswire) but from different sources/programs and epochs. All collected files appear on some list, with the exception of audio recordings that have been rejected during the earlier broadcast auditing process[7].

At the start of a work session the annotator logs into LDC's Annotation Workflow System (AWS) and chooses *Eval Data Selection* from their assigned list of tasks. AWS then launches the data selection toolkit and pre-loads the next unassigned file list from the pool of available data. After the tool launches, the annotator sees the list of files to review along with the first line of text in the file (where available) and the file size (in tokens or duration). The annotator clicks on each filename in turn, whereupon the tool displays the full text and/or audio for that file[8].

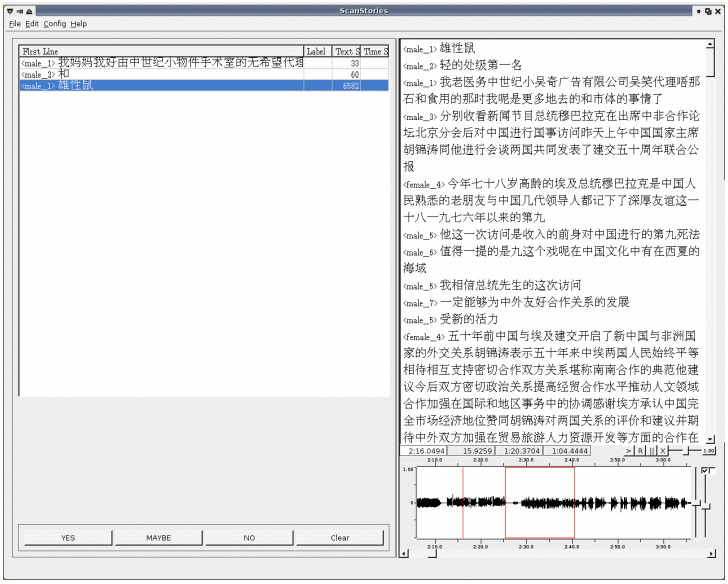

Figure 1: Audio and corresponding text display for selected story

The annotator reads through the text or ASR output, and/or listens to the audio recording, and makes an initial YES/MAYBE/NO judgment about the suitability of the file for inclusion in the selection pool based on the criteria specified in Section 2 above.

Because these criteria require segments of a particular length, annotators typically must designate one or more subsections of the file for selection. The toolkit includes functionality for marking off one or more segments within an

[6] The size of the file list depends on the genre.
[7] http://projects.ldc.upenn.edu/gale/task_specifications/ Audit_Procedure_Specificationv1.1.pdf
[8] For audio files, ASR text or closed captioning is displayed when available.

audio or text. When a segment is selected, the tool automatically updates the file inventory and file size (token or duration) information.

After the annotator labels a file or segment as YES or MAYBE, the tool launches a decision panel and the annotator makes the following judgments:

- Genre: NW, WL, NG, BN, BC, Other (with text box)
- Topic Category:
    1. Local or regional news - general
    2. National or international news - general
    3. Editorial or opinion
    4. Personal anecdote
    5. Human interest story
    6. Medical
    7. Legal
    8. Other
- Synopsis: 1-2 sentence description of content, in English
- ILR Rating: Less than 2; 2; 3; 3+
- Language/Dialect issues: text box
- Other comments: text box

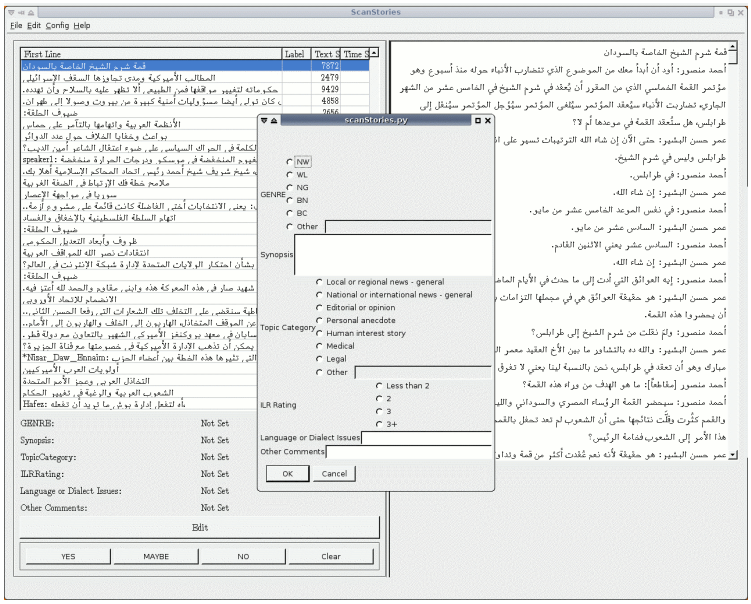

Figure 2: Completing judgments for selected story

After making the relevant judgments, the annotator moves on to the next file (or segment of the same file) for review.

At the close of the work session, the annotator is asked by AWS whether the file is complete or still in progress. If the file is labeled *complete*, AWS will allow the annotator to request another file for review or to log out. If the file is labeled *in progress*, AWS will launch the same file in the selection tool the next time the annotator logs in for a work session.

5 Automatic Scans on Selection Pool

Once the selection pool has been created using the manual process described above, NIST and LDC will run a series of automatic diagnostics to calculate log-perplexity and 3-gram hit-rate for documents in the pool. This will be used to identify documents that are outliers when compared to the rest of the selection pool and/or previous MT evaluation pools[9]. The process is as follows:

1. Compute the measure for all available documents in the pool.
2. Divide the desired distribution into a set of intervals - say 20.
3. Define a "current" distribution for the newly selected data (initially zero).
4. Take the difference in each bin between the desired distribution and the current distribution.
5. Find the interval where the difference is largest (i.e., the desired is larger than the current).
6. Select one of the documents in the pool in that interval.
7. Recompute the current probability distribution.
8. If we need more documents, go to step 4.

This process will rely on software developed by GALE sites[10].

6 Final selection of evaluation data set

Information available at the time of selecting the evaluation data sets include:

- Language
- Genre
- Source
- ILR
- Publication date
- Topic category
- Topic description
- Log perplexities
- Tri-gram hit rate
- Language / Dialect issues

[9] Documents identified as outliers will be flagged as such but will not necessarily be discarded from the evaluation data set.

[10] The current software donated by BBN handles text files only; so these scans cannot be run on audio data until transcripts have been created.

- Document section location {Begin, Middle, End}
- Document section word count (text)
- Document section duration (audio)
- * Possibly WER ratios for the audio data (score non-adaptive ASR vs. adaptive ASR)

As in past practices for data set creation, NIST will manually balance the test set selection across the above categories.

NIST will provide DARPA with comparisons between PHASE-1 data sets and PHASE-2 data sets, before the PHASE-2 evaluation data set is finalized.